# Estimation of Seasonal Components in Hourly Electricity Demand and Forecasting in West Bengal

**Author**
Jit Mondal

**Supervisor**
Dr. Raju Maiti
Assistant Professor, Economic Research Unit
**Indian Statistical Institute, Kolkata**

August 9, 2025

# Contents

# Chapter 1

# Introduction

## 1.1 Background

Electricity demand in any region is influenced by a combination of seasonal patterns, economic activities, and social behavior. In West Bengal, fluctuations in temperature, humidity, festival seasons, and industrial load create distinct patterns in electricity consumption. This project aims to analyze and estimate the different types of seasonal components (daily, weekly, and yearly) in hourly electricity demand and develop a forecasting model using statistical and machine learning approaches.

## 1.2 Objectives of the Study

This project aims to analyze and forecast hourly electricity demand in West Bengal by estimating seasonal components and evaluating predictive models. The specific objectives are:

1. **Seasonal Component Estimation:** Identify and estimate intraday, intraweek, and intrayear seasonal patterns in electricity demand.
2. **Time Series Decomposition:** Decompose the time series using:
   - Classical decomposition methods (e.g., additive/multiplicative).
   - Modern techniques like *MSTL* (Multiple Seasonal and Trend decomposition using Loess).
3. **Forecasting Model Development:** Develop and compare the following models:
   - *TBATS*

- Using interpretable Linear models.

4. **Model Evaluation:** Assess forecast accuracy using metrics such as:
   - Mean Squared Error (MSE).
   - Mean Absolute Error (MAE).
5. **Actionable Insights:** Provide recommendations for electricity providers to optimize load planning and demand management strategies.

# Chapter 2

# Methodology

## 2.1 Data Collection

The hourly electricity demand data for West Bengal was retrieved from NITI Aayog's *India Carbon & Energy Dashboard (ICED)* (https://iced.niti.gov.in/). The platform's interface required annual data extraction through the following workflow:

### 2.1.1 Step-by-Step Retrieval Process

1. **Access the Electricity Demand Module**:
   - Navigated to: *Electricity* → *Hourly Demand Curve*
2. **Configure Parameters**:
   - Set *Compare Between* dropdown to `State`
   - Selected *Year* (e.g., 2024 from the dropdown)
   - Choose *West Bengal* from the state selector
   - Click the *Download* icon (upper-right corner) for XLS export
3. **Repeat for All Years**:
   - Executed Steps 1–2 separately for 2017 through 2024
   - For 2024, only January–April data was available
4. **Consolidate the Data**
   - Temporal concatenation using Python's Pandas library
   - Conversion to CSV file

### 2.1.2 Data Characteristics

- **Format**: CSV files with columns `State`, `Date` and `Hourly Demand Met (in MW)`
- **Scope**: 1-Jan-2017 to 30-Apr-2024 (64248 hourly records)

The dataset exhibits complete coverage for all years except 2024 (partial year up to April). Special events like COVID-19 lockdowns (2020) and heatwaves (2022) are naturally embedded in the demand patterns.

# Chapter 3

# Exploratory Data Analysis

## 3.1 Exploratory Data Analysis of Daily Electricity Demand

### 3.1.1 Data Overview

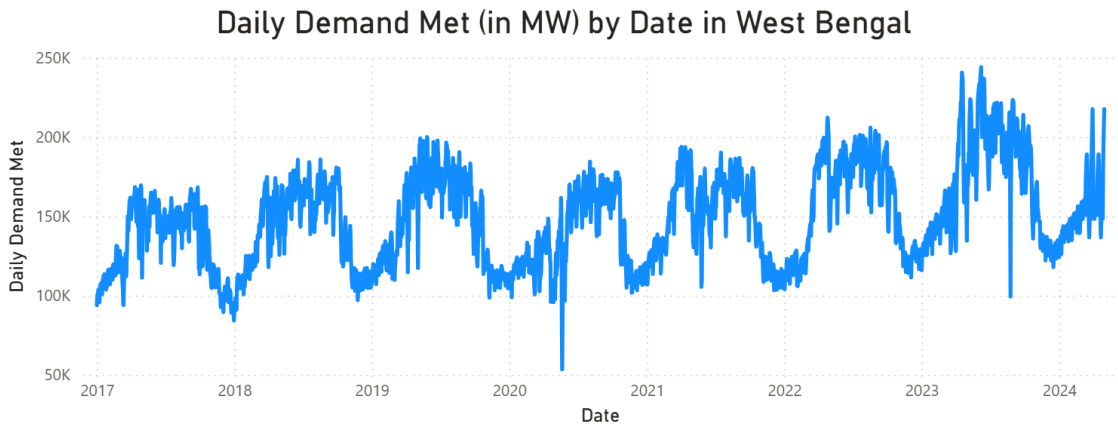


Figure 3.1: Daily electricity demand met (MW) in West Bengal (2017-2024)

The time series plot depicts **daily electricity demand (in MW)** for West Bengal from 2017 to 2024, revealing:

#### 1. Overall Trend

The time series plot of daily demand met (in MW) from 2017 to 2024 for West Bengal indicates a clear upward trend. The values typically start

around 100,000 MW in 2017 and reach over 200,000 MW by 2024. This consistent growth suggests an increase in electricity demand due to factors such as economic development, population growth, and greater electrification.

## 2. Seasonality

The data exhibits strong seasonal patterns, repeating annually. Each year, the demand typically peaks between the months of May and July, which likely corresponds to the summer season, when electricity consumption is high due to cooling appliances. Conversely, lower demands are observed around the winter months (December–January), when overall electricity usage tends to decrease.

## 3. Anomalies

There are noticeable sharp dips in certain years:

- Around mid-2020, which aligns with the COVID-19 lockdown period when industrial and commercial demand plummeted.
- Sporadic drops in other years could indicate data issues, power outages, natural disasters, or grid failures.
- A significant dip is also seen in mid-2023, which warrants deeper investigation.

## 4. Cyclic Behavior

Beyond seasonal effects, there appears to be multi-year cyclic behavior, with demand showing accelerated increases every few years. This could reflect infrastructure improvements, new policies, or large-scale industrial developments.

## 5. Volatility

The variability in the daily demand has increased over time. The earlier years (2017–2019) show more stable demand levels, while later years (2021 onward) exhibit larger fluctuations. This could be due to changes in

industrial usage patterns, weather variability, or increased sensitivity of the grid.

### 6. Forecasting Potential

Given the clear trend and pronounced seasonality, this time series is suitable for various forecasting techniques, such as:

- MSTL (Multiple Seasonal-Trend Decomposition)
- TBATS
- Models with dummy variables (e.g., weekday/weekend, month indicators)

## 3.2 Seasonal Hourly Demand Patterns

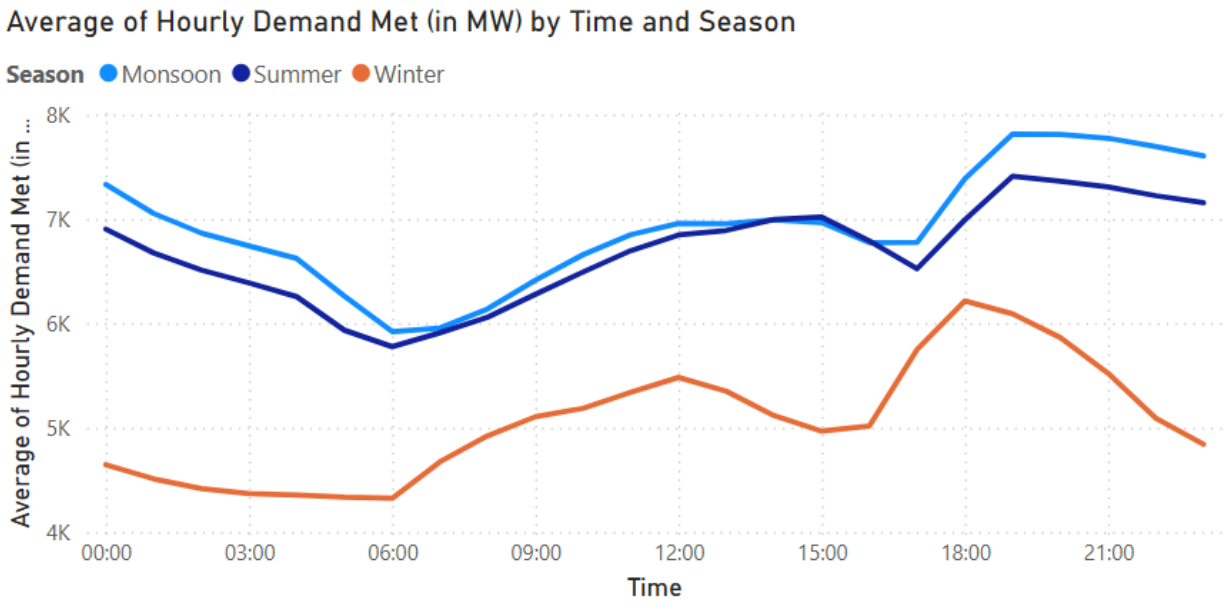


Figure 3.2: Hourly demand variations across seasons in West Bengal showing characteristic troughs and peaks

The plot showing the average hourly demand met (in MW) by time and season offers critical insights into intraday electricity consumption behavior across different climatic conditions.

### 3.2.1 Summer Season (March–June)

- Summer shows the highest demand throughout the day.
- Demand drops to its minimum around 06:00 AM (below 6,000 MW), then increases steadily.
- A peak is observed between 18:00 and 20:00 hours, reaching close to 8,000 MW, likely due to high evening residential usage (air conditioning, lighting, etc.).

### 3.2.2 Monsoon Season (July–October)

- Monsoon follows a similar intraday pattern as summer but with slightly lower magnitude.
- The morning dip and evening rise are present but more subdued.
- The peak is also seen around 19:00–20:00 hours, approaching 7,500 MW.

### 3.2.3 Winter Season (November–February)

- Winter exhibits the lowest overall demand among the three seasons.
- The demand curve is relatively flatter throughout the day.
- An unusual evening peak is observed around 18:00 hours, suggesting usage related to heating, lighting, and evening activities.
- Even at its peak, the demand remains below 6,500 MW.

### 3.2.4 Overall Observations

- All seasons show a typical ”U-shaped” daily pattern — demand is lowest in early morning (04:00–06:00) and highest in the evening.
- The amplitude of variation is highest in summer and lowest in winter.
- The consistent evening peak across seasons implies a strong behavioral component in power usage.

### 3.2.5 Implications for Load Forecasting

- These hourly patterns emphasize the need for incorporating time-of-day and seasonality effects into forecasting models.
- Time-dependent dummy variables or fourier series (sine/cosine terms) may help capture these cyclical behaviors.
- Real-time load forecasting should consider seasonally adaptive modeling to adjust to changing daily usage curves.

## 3.3 Day-Night Electricity Demand Analysis

The figure below illustrates the **average hourly demand met (in MW)** by date, separated by **daytime** and **nighttime** periods in West Bengal over the span of 2017 to 2024.

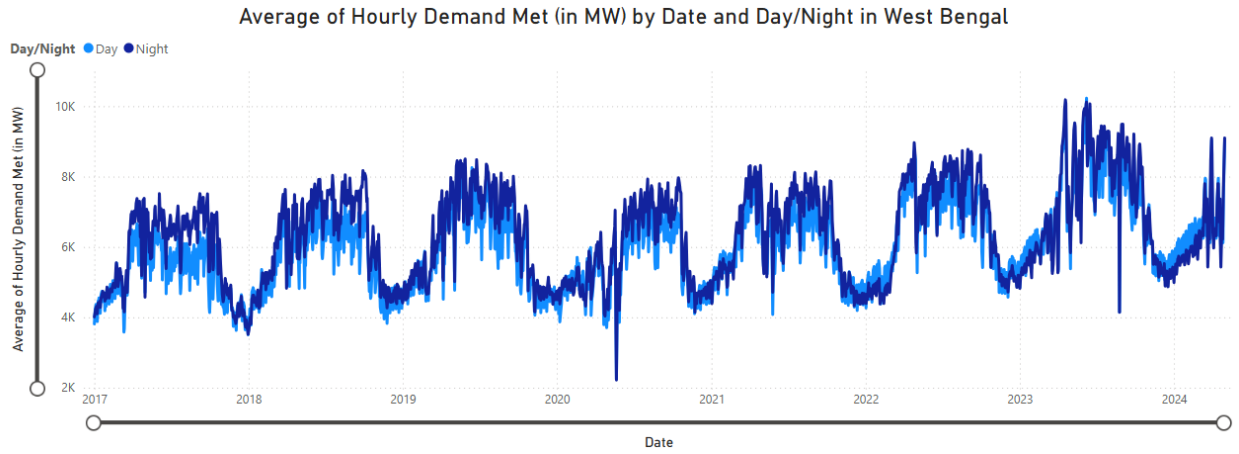


Figure 3.3: Day vs. night electricity demand patterns in West Bengal (2017-2024)
*Day: 06:00-18:00, Night: 18:00-06:00*

- **Definition:**
  - **Day:** 06:00 AM to 06:00 PM
  - **Night:** 06:00 PM to 06:00 AM

Table 3.1: Day-Night Demand Comparison

| Metric | Day | Night |
|---|---|---|
| Average Demand (MW) | 6,073.85 | 6,359.54 |
| Median Demand (MW) | 6,028.66 | 6,377.69 |
| Demand Difference | +285.69 MW | |
| Relative Difference | +4.7% higher at night | |

### 3.3.1 Key Statistics

### 3.3.2 Key Observations

- **Higher Demand at Night:**
  - Nighttime demand is consistently higher than daytime demand across all years.
  - This can be attributed to residential lighting and cooling appliance usage during evening and night hours.
- **Strong Seasonal Variation:**
  - Both day and night series show clear seasonal patterns.
  - Peaks are observed during the summer months (March–June), particularly at night, due to increased use of fans, coolers, and air conditioners.
  - Dips are consistently seen during winter periods (November–February), where electricity demand is generally lower.
- **COVID-19 Impact:**
  - In 2020, there is a sharp and sudden drop in both day and night demand — particularly visible around March–May, coinciding with nationwide lockdowns.
- **Post-2020 Recovery and Surges:**
  - Demand has shown a sharp rise post-pandemic, with some of the highest peaks occurring during the summer of 2023.

  - Nighttime peaks appear to grow more pronounced in recent years, indicating increasing nighttime consumption trends.

- **Intra-Annual Trends:**
  - Regular yearly demand surges and drops suggest strong influence of seasonal temperatures on electricity usage.
  - This supports the inclusion of seasonal dummy variables in modeling approaches.

### 3.3.3 Modeling Implications

- Day vs Night behavior must be explicitly modeled for accurate forecasts, especially when forecasting at hourly/daily granularity.
- The clear separation between day and night demand supports the use of a categorical time-of-day feature in time series decomposition and regression models.
- Increases in nighttime consumption suggest future electricity infrastructure planning may need to focus more on evening peak load management.

## 3.4 Exploratory Data Analysis of Quarterly Data

### Overview

The chart titled *"Average of Hourly Demand Met (in MW) by Year and Quarter for West Bengal"* presents quarterly electricity demand data from 2017 to 2024.

### Variable Information

- **Dependent Variable:** Average Hourly Demand Met (in MW)
- **Independent Variables:** Year and Quarter

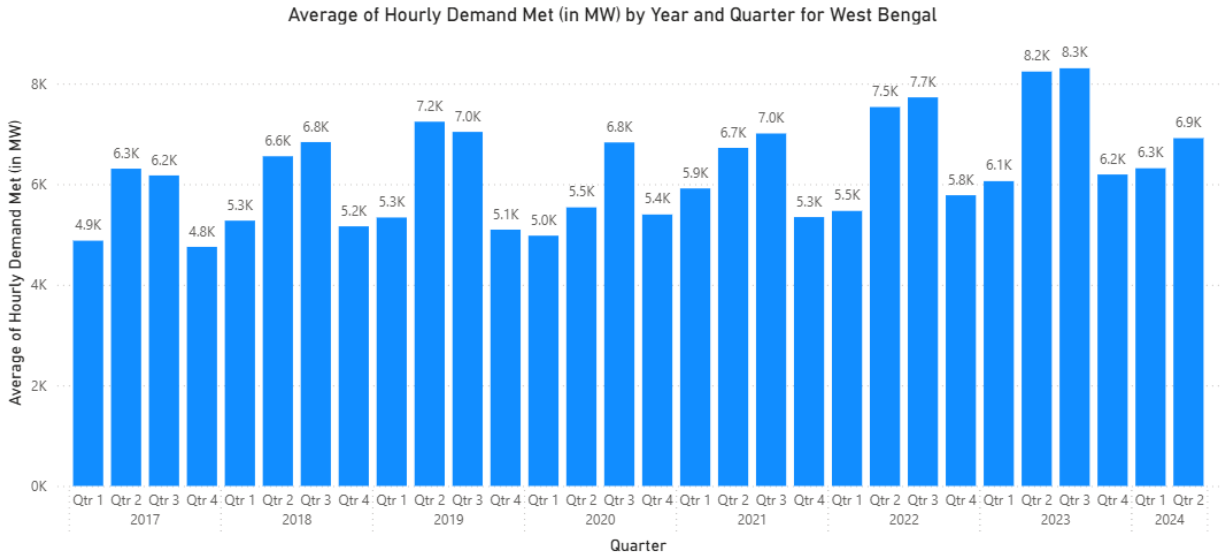


Figure 3.4: Quarterly electricity demand patterns in West Bengal (2017-2024)

## Key Observations

1. **Overall Trend:** There is a general upward trend in electricity demand from 2017 to 2023, with a peak in Q3 2023 (8.3K MW).

2. **Seasonality:** Higher demand is observed during Q2 and Q3 (summer months), indicating seasonal variation likely due to cooling requirements.

3. **COVID-19 Impact:** A noticeable dip occurs in Q1 and Q2 of 2020, coinciding with the nationwide lockdown during the pandemic.

4. **Post-Pandemic Recovery:** From 2021 onwards, demand rebounds strongly, with 2022 and 2023 showing record highs.

5. **Recent Trends:** In 2024, demand seems to stabilize or decline slightly (Q1: 6.3K MW, Q2: 6.9K MW).

## Interpretation

The data suggests a steady increase in electricity usage in West Bengal over the years. Seasonal patterns are evident, with peak demands during the

summer quarters. The sharp drop in 2020 aligns with the pandemic's economic disruption. Post-2020, the recovery indicates a possible structural shift in the demand baseline, potentially due to increased electrification, infrastructure development, and economic recovery.

## Remarks

The increasing trend, combined with quarterly seasonality, can help in predictive modeling for electricity demand. Such insights are valuable for power generation planning and infrastructure management.

# Chapter 4

# Some Theoretical Background

## 4.1 Multiple Seasonal-Trend Decomposition using Loess (MSTL)

MSTL is an extension of the classic STL (Seasonal-Trend decomposition using Loess) method, designed to handle time series data with multiple seasonal patterns. Unlike STL, which can only decompose one seasonal component, MSTL generalizes the model to multiple seasonalities (e.g., daily and yearly cycles in electricity demand data).

### 4.1.1 Mathematical Background

Let $y_t$ be the observed time series. MSTL models the series as:

$$y_t = T_t + \sum_{i=1}^{K} S_t^{(i)} + R_t$$

where:

- $T_t$ is the smooth trend component
- $S_t^{(i)}$ is the $i$-th seasonal component with period $p_i$
- $R_t$ is the residual (remainder) component

The decomposition is performed in a **stepwise additive** manner:

$$y_t \xrightarrow{\text{remove } S^{(1)}} y_t^{(1)} \xrightarrow{\text{remove } S^{(2)}} \cdots \xrightarrow{\text{remove } S^{(K)}} y_t^{(K)} \xrightarrow{\text{estimate trend}} T_t$$

Each seasonal component and the trend are extracted using **LOESS smoothing**, a non-parametric local regression technique.

### 4.1.2 Lowess (LOESS) Smoothing

LOESS (Locally Estimated Scatterplot Smoothing) fits a low-degree polynomial (typically linear or quadratic) to localized subsets of data.

For a given point $x_0$, LOESS performs weighted least squares:

$$\hat{\beta}(x_0) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} w_i(x_0)(y_i - \beta_0 - \beta_1(x_i - x_0))^2$$

where $w_i(x_0)$ is a weight function, usually a tricube kernel:

$$w_i(x_0) = \left(1 - \left|\frac{x_i - x_0}{d(x_0)}\right|^3\right)^3$$

and $d(x_0)$ is the distance to the furthest neighbor within the local span.

### 4.1.3 Steps of MSTL

1. For each seasonal period $m_j$, estimate seasonal component $S_{j,t}$ using STL on residuals from the previous step.
2. Subtract all seasonal components from $Y_t$ to obtain a detrended, deseasonalized series.
3. Apply LOESS to the residual to extract trend $T_t$.
4. Remainder $R_t = Y_t - T_t - \sum_j S_{j,t}$.

### 4.1.4 Illustrative Example

Let us consider a simple synthetic time series:

$$Y = \{15, 18, 21, 20, 25, 28, 33, 30, 25, 22, 18, 15\}$$

Suppose the time series has monthly seasonality ($m = 12$). We first remove seasonality by averaging each month across cycles. With just one cycle, assume centered moving average as estimate:

$$\hat{S}_t = \text{Centered 12-month moving average}$$

Let us compute trend using LOESS on:

$$Y_t - \hat{S}_t \Rightarrow \text{LOESS smoothing}$$

Then:

$$R_t = Y_t - \hat{S}_t - \hat{T}_t$$

For demonstration, suppose we compute:

$$\hat{S}_t = \{0, 1, 2, 1, 2, 3, 2, 1, 0, -1, -2, -1\}$$

$$\hat{T}_t = \{15, 17, 19, 19, 23, 25, 31, 29, 25, 23, 20, 16\}$$

Then,

$$R_t = Y_t - \hat{S}_t - \hat{T}_t$$

E.g., for $t = 2$: $Y_2 = 18$, $\hat{S}_2 = 1$, $\hat{T}_2 = 17$, $\Rightarrow R_2 = 0$.

### 4.1.5 Conclusion

MSTL is a powerful decomposition tool, particularly suited for time series with multiple seasonalities. It uses robust LOESS smoothing to extract trend and seasonal components iteratively, allowing clearer residual analysis and improved forecasting.

The decomposition results in three components:

- Trend: captures long-term changes (here, a constant level of 10)
- Daily seasonality: repeats every 24 hours
- Weekly seasonality: repeats every 168 hours
- Residual: the remaining noise

### 4.1.6 Applications of MSTL

MSTL is particularly suited for high-frequency time series such as:

- Electricity demand (daily, weekly, yearly cycles)
- Website traffic or server loads (hourly, daily patterns)
- Environmental sensor data

By isolating each component, MSTL enables cleaner modeling, forecasting, and anomaly detection (e.g., by modeling the residuals separately using ARIMA).

## 4.2 TBATS Model

The TBATS (Trigonometric, Box-Cox transform, ARMA errors, Trend, and Seasonality) model is designed to handle complex seasonal patterns in electricity demand time series. The model decomposes as follows:

### 4.2.1 Core Components

$$y_t^{(\omega)} = \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t + \epsilon_t \tag{4.1}$$

where:

- $y_t^{(\omega)}$: Box-Cox transformed demand at time $t$ with parameter $\omega$
- $\ell_t$: Local level
- $b_t$: Trend component
- $\phi$: Damped trend parameter ($0 < \phi \leq 1$)
- $s_t^{(i)}$: Seasonal component for period $m_i$
- $d_t$: ARMA(p,q) process for residuals
- $\epsilon_t \sim N(0, \sigma^2)$: Gaussian noise

### 4.2.2 Component Details

**Box-Cox Transformation**

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{\omega}-1}{\omega} & \omega \neq 0 \\ \log y_t & \omega = 0 \end{cases} \tag{4.2}$$

- Stabilizes variance in demand fluctuations
- $\omega$ optimized via profile likelihood

**Trend Component**

$$\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha \epsilon_t \tag{4.3}$$
$$b_t = \phi b_{t-1} + \beta \epsilon_t \tag{4.4}$$

where $\alpha, \beta$ are smoothing parameters.

**Trigonometric Seasonality**

For each seasonality $k$ with period $m_k$:

$$s_t^{(k)} = \sum_{j=1}^{L_k} s_{j,t}^{(k)} \tag{4.5}$$
$$s_{j,t}^{(k)} = s_{j,t-1}^{(k)} \cos(\lambda_k j) + s_{j,t-1}^{*(k)} \sin(\lambda_k j) + \gamma_1^{(k)} \epsilon_t \tag{4.6}$$
$$s_{j,t}^{*(k)} = -s_{j,t-1}^{(k)} \sin(\lambda_k j) + s_{j,t-1}^{*(k)} \cos(\lambda_k j) + \gamma_2^{(k)} \epsilon_t \tag{4.7}$$

where $\lambda_k = 2\pi/m_k$ and $L_k$ is the number of harmonics.

**ARMA Residuals**

$$d_t = \sum_{i=1}^{p} \psi_i d_{t-i} + \sum_{j=1}^{q} \theta_j \eta_{t-j} + \eta_t \tag{4.8}$$

where $\eta_t \sim N(0, \sigma_\eta^2)$.

### 4.2.3 Parameter Estimation

The model estimates parameters via:

$$\Theta = \{\omega, \phi, \alpha, \beta, \{\gamma_1^{(k)}, \gamma_2^{(k)}\}_{k=1}^{T}, \{\psi_i\}_{i=1}^{p}, \{\theta_j\}_{j=1}^{q}\} \tag{4.9}$$

using maximum likelihood estimation:

$$\hat{\Theta} = argmax_{\Theta}\mathcal{L}(\Theta|y_1, \ldots, y_n) \tag{4.10}$$

### 4.2.4 Advantages for Electricity Demand

- Handles **multiple seasonalities** (daily, weekly, annual)
- Accommodates **non-integer periods** (e.g., 24.2-hour daily cycle)
- Robust to **missing data** common in power systems
- Automatic **harmonic selection** via $L_k$ optimization

## 4.3 SARIMAX Model

This section presents a step-by-step mathematical derivation and background starting from ARMA models, extending to ARIMA, then SARIMA, and finally SARIMAX. We discuss stationarity/invertibility conditions, estimation (MLE / Kalman filter), forecasting equations, and diagnostic checks.

### 4.3.1 Preliminaries: Backshift and Difference Operators

Define the backshift (lag) operator $B$ by

$$By_t = y_{t-1}, \qquad B^k y_t = y_{t-k}.$$

The (non-seasonal) first difference operator is $\nabla = 1 - B$, and the seasonal difference of period $s$ is $\nabla_s = 1 - B^s$.

### 4.3.2 ARMA$(p, q)$

An ARMA$(p, q)$ process $\{y_t\}$ satisfies

$$\phi(B)y_t \;=\; \theta(B)\varepsilon_t, \tag{4.11}$$

where $\{\varepsilon_t\}$ is white noise with $E[\varepsilon_t] = 0$ and $\mathrm{Var}(\varepsilon_t) = \sigma^2$, and

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \qquad \theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

**Stationarity.** The ARMA process is (weakly) stationary if the roots of $\phi(z) = 0$ lie *outside* the complex unit circle ($|z| > 1$). Equivalently, $\phi(B)$ is invertible as a power series:

$$y_t = \phi(B)^{-1}\theta(B)\varepsilon_t = \psi(B)\varepsilon_t,$$

with $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_j |\psi_j| < \infty$.

**Invertibility.** The MA polynomial $\theta(z) = 0$ should have roots outside the unit circle to express $\varepsilon_t$ as an infinite AR series, which is useful for identification and estimation.

### 4.3.3 ARIMA$(p, d, q)$

If $\{y_t\}$ is non-stationary, differencing of order $d$ may be applied:

$$\phi(B)(1 - B)^d y_t = \theta(B)\varepsilon_t.$$

Setting $x_t = (1 - B)^d y_t$ yields the stationary ARMA$(p, q)$ model for $x_t$.

### 4.3.4 Seasonal Extension: SARIMA$(p, d, q) \times (P, D, Q)_s$

To incorporate seasonality of period $s$, include seasonal polynomials:

$$\Phi(B^s)\,\phi(B)\,(1 - B)^d(1 - B^s)^D y_t \;=\; \Theta(B^s)\,\theta(B)\,\varepsilon_t, \tag{4.12}$$

where

$$\Phi(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps}, \qquad \Theta(B^s) = 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs}.$$

This models both non-seasonal dynamics (via $\phi, \theta$) and seasonal dynamics (via $\Phi, \Theta$), with non-seasonal differencing $d$ and seasonal differencing $D$.

### 4.3.5 Inclusion of Exogenous Regressors: SARIMAX

Include a vector of exogenous regressors $X_t$ (dimension $k$) with coefficient vector $\beta$:

$$\Phi(B^s)\,\phi(B)\,(1-B)^d(1-B^s)^D y_t \;=\; \Theta(B^s)\,\theta(B)\,\varepsilon_t \;+\; \beta^\top X_t. \tag{4.13}$$

Equivalently, after applying the combined difference operator $\Delta(B) = (1-B)^d(1-B^s)^D$, we can write

$$\tilde{\phi}(B)\tilde{y}_t = \tilde{\theta}(B)\varepsilon_t + \beta^\top X_t, \qquad \tilde{y}_t = \Delta(B)y_t,\ \tilde{\phi} = \Phi\phi,\ \tilde{\theta} = \Theta\theta.$$

### 4.3.6 State-Space Representation and Kalman Filter

SARIMAX can be written in state-space form (useful for MLE via Kalman filter):

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{F}\mathbf{x}_t + \mathbf{G}\varepsilon_t,\\ y_t &= \mathbf{H}^\top\mathbf{x}_t + \beta^\top X_t + \varepsilon_t,\end{aligned}$$

with appropriate companion matrices $\mathbf{F}, \mathbf{G}, \mathbf{H}$. The Kalman filter computes the one-step-ahead prediction $\hat{y}_{t|t-1}$ and the prediction error $\nu_t = y_t - \hat{y}_{t|t-1}$, yielding the Gaussian log-likelihood:

$$\ell(\theta) = -\frac{1}{2}\sum_{t=1}^{T}\left(\ln|\mathbf{S}_t| + \nu_t^\top\mathbf{S}_t^{-1}\nu_t\right) - \frac{nT}{2}\ln(2\pi),$$

where $\mathbf{S}_t$ is the prediction error variance. Maximizing $\ell(\theta)$ gives MLEs of parameters $\theta = (\phi, \theta, \Phi, \Theta, \beta, \sigma^2)$.

### 4.3.7 Parameter Estimation

- **Maximum Likelihood:** MLE via numerical optimization of the Gaussian log-likelihood (often implemented with Kalman filter for state-space).
- **Conditional Sum-of-Squares / Yule–Walker:** Initial estimates may be obtained by simpler methods and refined by MLE.

### 4.3.8 Forecasting Equations

Let $\hat{y}_{t+h|t}$ denote the $h$-step-ahead forecast given information up to time $t$. For an ARMA-type representation

$$y_t = \sum_{i=1}^{\infty} \psi_i \varepsilon_{t-i},$$

the $h$-step forecast is computed recursively using the model and replacing future errors by their expectation (zero). In practice:

$$\hat{y}_{t+1|t} = \phi_1 \hat{y}_{t|t} + \cdots + \phi_p \hat{y}_{t+1-p|t} + \beta^\top X_{t+1}$$

and for $h > 1$ repeat recursively, plugging forecasts into the AR terms and zeros into future innovations. If the model is written in companion matrix form, the $h$-step forecast can be expressed as

$$\hat{\mathbf{x}}_{t+h|t} = \mathbf{F}^h \hat{\mathbf{x}}_{t|t} + \sum_{j=0}^{h-1} \mathbf{F}^j \mathbf{G}\, E[\varepsilon_{t+h-j}] = \mathbf{F}^h \hat{\mathbf{x}}_{t|t},$$

and $\hat{y}_{t+h|t} = \mathbf{H}^\top \hat{\mathbf{x}}_{t+h|t} + \beta^\top X_{t+h}$.

### 4.3.9 Statistical Properties: Stationarity and Invertibility Conditions

- **Stationarity:** All roots of $\phi(z)\Phi(z^s) = 0$ must lie outside the unit circle.
- **Invertibility:** All roots of $\theta(z)\Theta(z^s) = 0$ must lie outside the unit circle.

These conditions ensure finite-order representations as convergent infinite series and stable forecasting.

### 4.3.10 Model Selection and Diagnostics

**Information Criteria:** Choose orders $(p, d, q, P, D, Q)$ by minimizing AIC or BIC:

$$\text{AIC} = -2\ell(\hat{\theta}) + 2k, \qquad \text{BIC} = -2\ell(\hat{\theta}) + k \ln(T),$$

where $k$ is the number of estimated parameters.

**Residual Diagnostics:** After fitting, analyze standardized residuals $\hat{\varepsilon}_t$:

- **Ljung–Box test:** test $H_0$ that residuals are uncorrelated up to lag $m$:

$$Q(m) = T(T+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{T-k} \sim \chi^2_{m-p-q-P-Q},$$

  where $\hat{\rho}_k$ are sample autocorrelations of residuals.

- **Normality:** Jarque–Bera test on residuals for Gaussianity.

- **Heteroskedasticity:** Tests (e.g., Engle's ARCH test) to detect conditional variance.

### 4.3.11 Interpretation of Coefficients (Practical Notes)

- AR coefficients near 1 indicate high persistence; seasonal AR near 1 indicate strong seasonal persistence.

- MA coefficients capture the influence of recent shocks; negative seasonal MA can indicate alternating seasonal corrections.

- Exogenous $\beta$ coefficients quantify the direct, contemporaneous effect of external variables $X_t$.

### 4.3.12 When SARIMAX May Fail or Underperform

- Missing relevant exogenous variables (omitted variable bias) can leave structured residuals.

- Nonlinear relationships or regime changes (structural breaks) are not captured by linear SARIMAX.

- Heteroskedasticity and heavy-tailed errors violate Gaussian assumptions; consider GARCH or robust methods.

### 4.3.13 Summary of Practical Steps for Fitting SARIMAX

1. Inspect ACF/PACF and domain knowledge to propose $(p, d, q, P, D, Q, s)$.

2. Difference the series ($d$, $D$) to achieve stationarity if necessary (ADF test).

3. Optionally include exogenous regressors $X_t$ (lagged or contemporaneous).

4. Fit by MLE (Kalman filter) and inspect AIC/BIC for model comparison.

5. Validate with residual diagnostics (Ljung–Box, JB, ARCH tests).

6. Produce h-step forecasts using recursive equations and provide forecast intervals using the prediction variance from the Kalman filter.

# Chapter 5

# Data Analysis

## 5.1 Autocorrelation Analysis

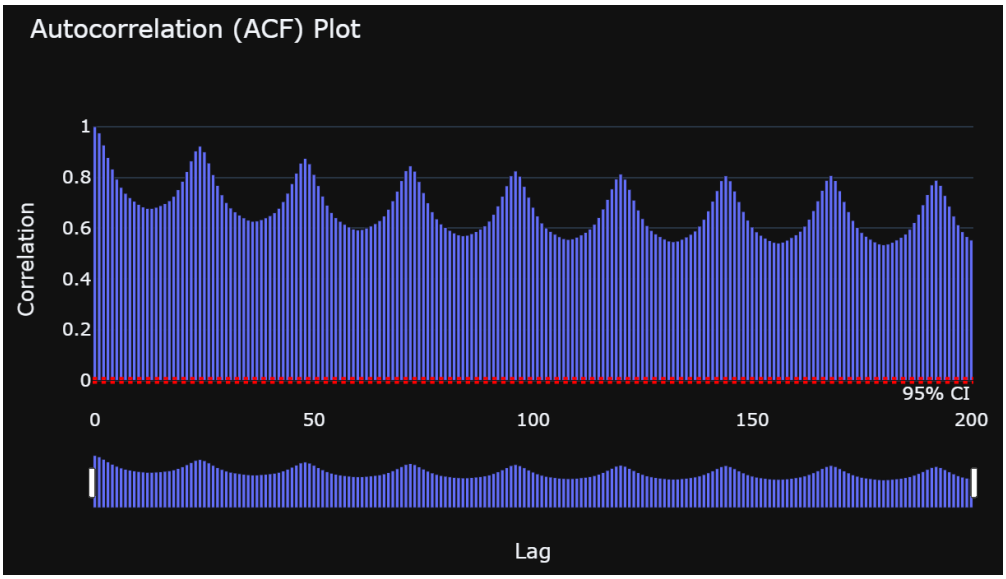


Figure 5.1: Autocorrelation Function (ACF) plot of hourly electricity demand in West Bengal

To further investigate the temporal dependence in the series, we examine the **Autocorrelation Function (ACF)** plot of the hourly demand met:

- **Significant Lag-24 Spikes:** The ACF plot reveals prominent spikes at lags that are multiples of 24 (i.e., 24, 48, 72, ...). This is a strong indication of a **daily seasonality** in the data.

- **Gradual Decay:** The autocorrelation decreases slowly across lags, suggesting a persistent correlation structure—typical for non-stationary time series.
- **Statistical Significance:** All major spikes lie beyond the 95% confidence interval bounds (marked in red), confirming statistical significance.
- **Implication:** The daily seasonality reflects a regular pattern of energy consumption that repeats every 24 hours, driven by human activity cycles.

**Remarks:** The data exhibits clear signs of **non-stationarity, strong daily seasonality**, and **distinct day-night consumption patterns**, which should be accounted for in any subsequent modeling or forecasting efforts.

## 5.2 Decomposition Results using MSTL

To analyze the underlying structure of the hourly electricity demand data, we applied **MSTL (Multiple Seasonal-Trend decomposition using LOESS)**. Based on insights from the data:

- The **first seasonal period** was chosen as 24 (hours) since the **Autocorrelation Function (ACF)** plot revealed strong correlation at lag 24, indicating **daily seasonality**.
- The **second seasonal period** was selected as 8760 (hours) based on domain knowledge and exploratory analysis, which showed a clear **yearly seasonal pattern**.

The MSTL model was run with `periods = [24, 8760]`. The decomposition yielded four components:

- The overall **Trend**,
- **Daily Seasonality** (24-hour cycle),
- **Yearly Seasonality** (8760-hour cycle),

- The remaining **Residuals**.

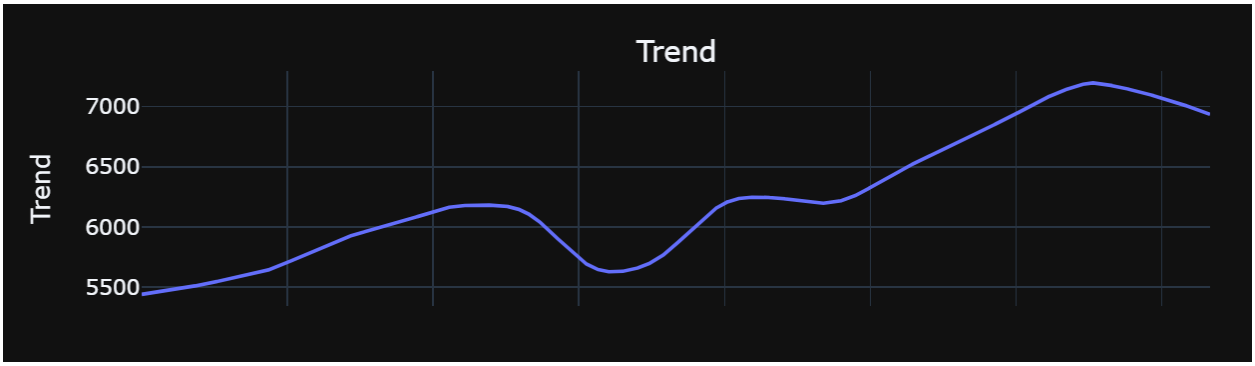


Figure 5.2: Trend Component extracted by MSTL

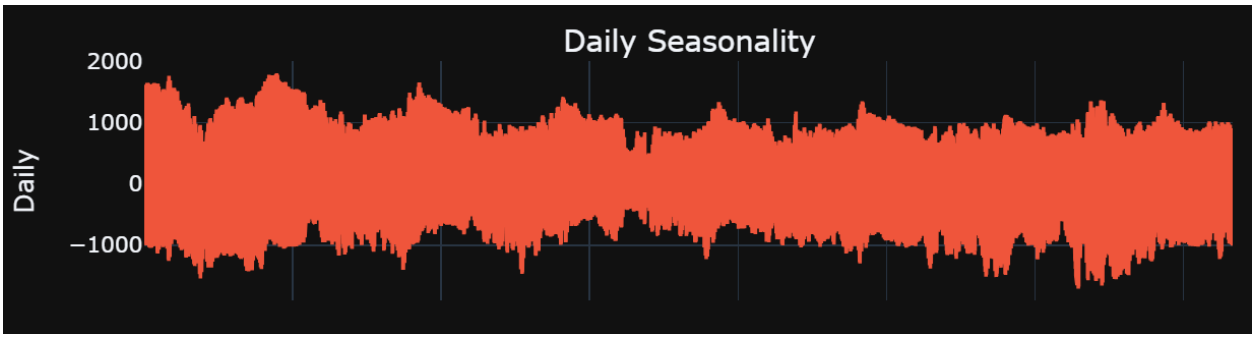


Figure 5.3: Daily Seasonality Component (24-hour periodicity)

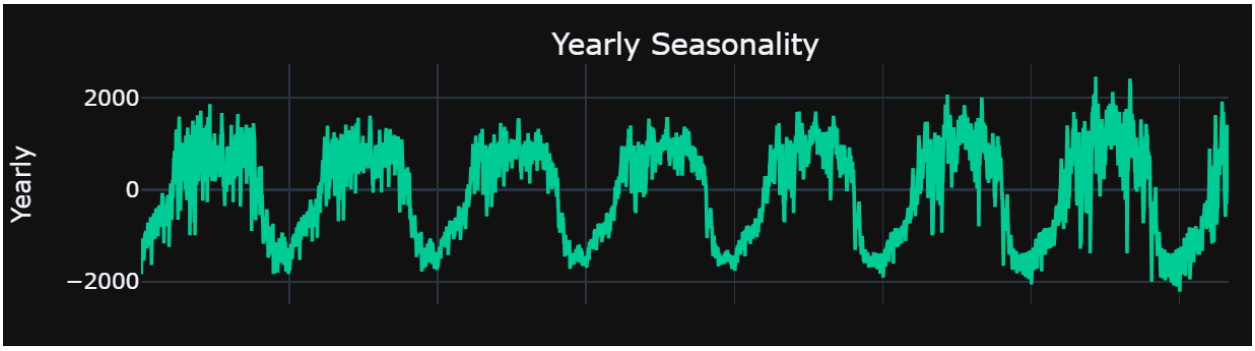


Figure 5.4: Yearly Seasonality Component (8766-hour periodicity)

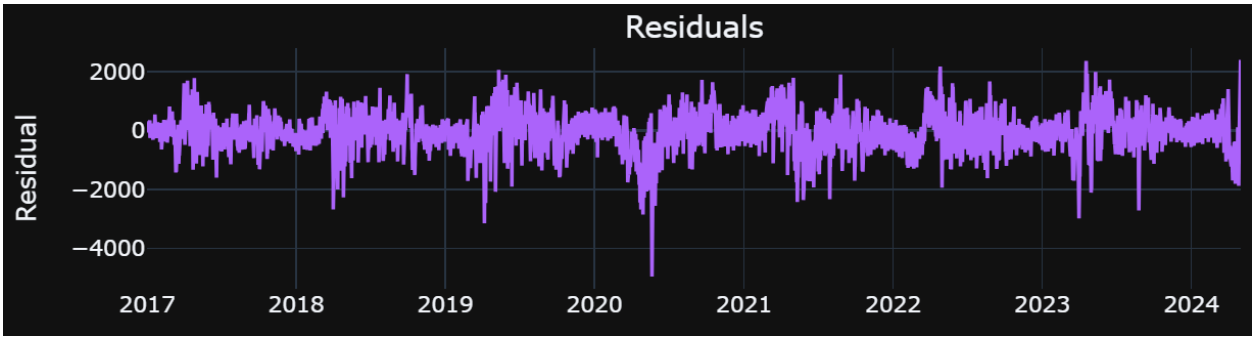


Figure 5.5: Residual Component after removing trend and seasonal effects

Each component was extracted using robust LOESS smoothing, enabling the isolation of distinct patterns across multiple seasonalities.

### 5.2.1 ACF Analysis of Residuals

**MSTL Decomposition Residuals**

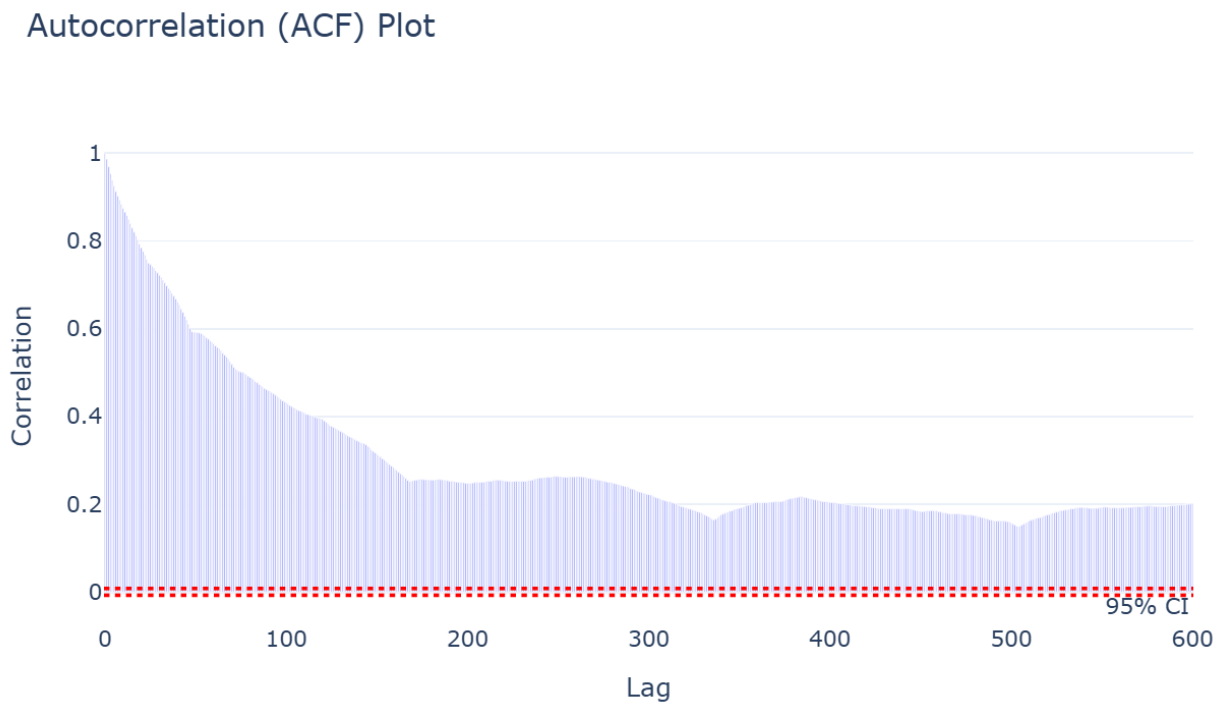


Figure 5.6: ACF of the MSTL Residual

The Autocorrelation Function (ACF) plot of the residuals from the MSTL decomposition (Figure 5.6) shows no significant spikes, suggesting that the

seasonal and trend components have been effectively captured by the MSTL model. This indicates that the residuals are approximately white noise and there is minimal autocorrelation left in the data. This is a positive sign of a well-fitted decomposition model.

### 5.2.2 ADF Test on MSTL Residuals

To assess the stationarity of the residuals resulting from the MSTL decomposition, we applied the Augmented Dickey-Fuller (ADF) test. This test evaluates the null hypothesis that the time series has a unit root (i.e., it is non-stationary) against the alternative hypothesis of stationarity.

- **ADF Test Statistic:** `-15.4589`
- **p-value:** $2.72 \times 10^{-28}$
- **Used lags:** 61
- **Number of observations:** 64186
- **Critical Values:**
  - 1% level: `-3.4305`
  - 5% level: `-2.8616`
  - 10% level: `-2.5668`

**Interpretation:** Since the ADF test statistic (`-15.4589`) is far less than all the critical values at 1%, 5%, and 10% significance levels, and the p-value is effectively zero ($2.72 \times 10^{-28}$), we reject the null hypothesis of a unit root. Therefore, we conclude that the residuals are **stationary**, indicating that the MSTL decomposition has successfully removed the trend and seasonality components from the time series.

## 5.3 Modeling and Forecasting

The data was split as follows:

- **Training period:** January 2017 to December 2022
- **Forecasting period:** January 2023 to April 2024

### 5.3.1 Trend Component

After extracting the trend component using MSTL decomposition, we aimed to forecast its future values using a smooth, flexible model. For this, we employed a **natural cubic spline** with:

- **Polynomial degree:** 3 (cubic spline),
- **Degrees of freedom (df):** 200 (Based on AIC).

This spline allows us to capture non-linear long-term movements in electricity demand without overfitting short-term noise.

Using the spline model trained on the historical trend, we generated predictions for the trend component during the forecasting window. The forecasted trend provides a smoothed approximation of the long-term change in electricity demand and serves as a crucial component in the full demand prediction.

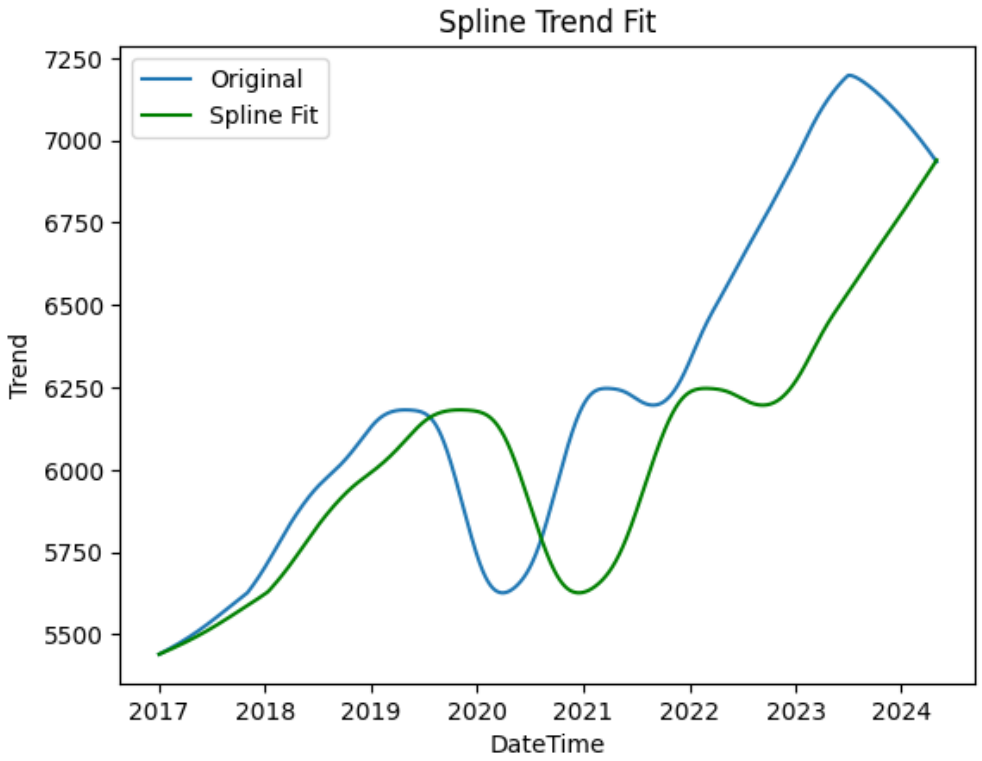


Figure 5.7: Trend Component Forecast using Natural Cubic Spline (df=200)

### 5.3.2 Modeling Daily Seasonality: Dummy Variables vs TBATS

To model the **daily seasonality** component (period = 24 hours) extracted via MSTL, two different approaches were considered:

#### 1. Dummy Variable Encoding

We created **hour-of-day dummy variables** to model the daily pattern. This approach assumes each hour has a distinct effect on electricity demand. It is straightforward and interpretable, making it useful for capturing fixed hourly patterns across days.

#### 2. TBATS Model

We also used the **TBATS** model, which can handle multiple seasonalities, non-linear trends, and complex autocorrelations. Though TBATS is typically used for long-period seasonality, it was applied here to assess whether it could model the daily pattern more flexibly than dummy encoding.

To evaluate both methods, we trained models on data from **2017 to 2022** and predicted daily seasonality over the period **January 2023 to April 2024**. The Mean Squared Errors (MSE) of the two methods are presented below.

Table 5.1: Comparison of Daily Seasonality Modeling Methods

| **Modeling Approach** | **MSE** | **Remarks** |
|---|---|---|
| Dummy Variables | 164654.01 | Lower error |
| TBATS | 164713.37 | Slightly higher error |

While both models achieved very similar accuracy, the dummy variable approach slightly outperformed TBATS in terms of MSE. Additionally, the interpretability and simplicity of the dummy variable model make it more suitable for capturing fixed hourly effects in this context.

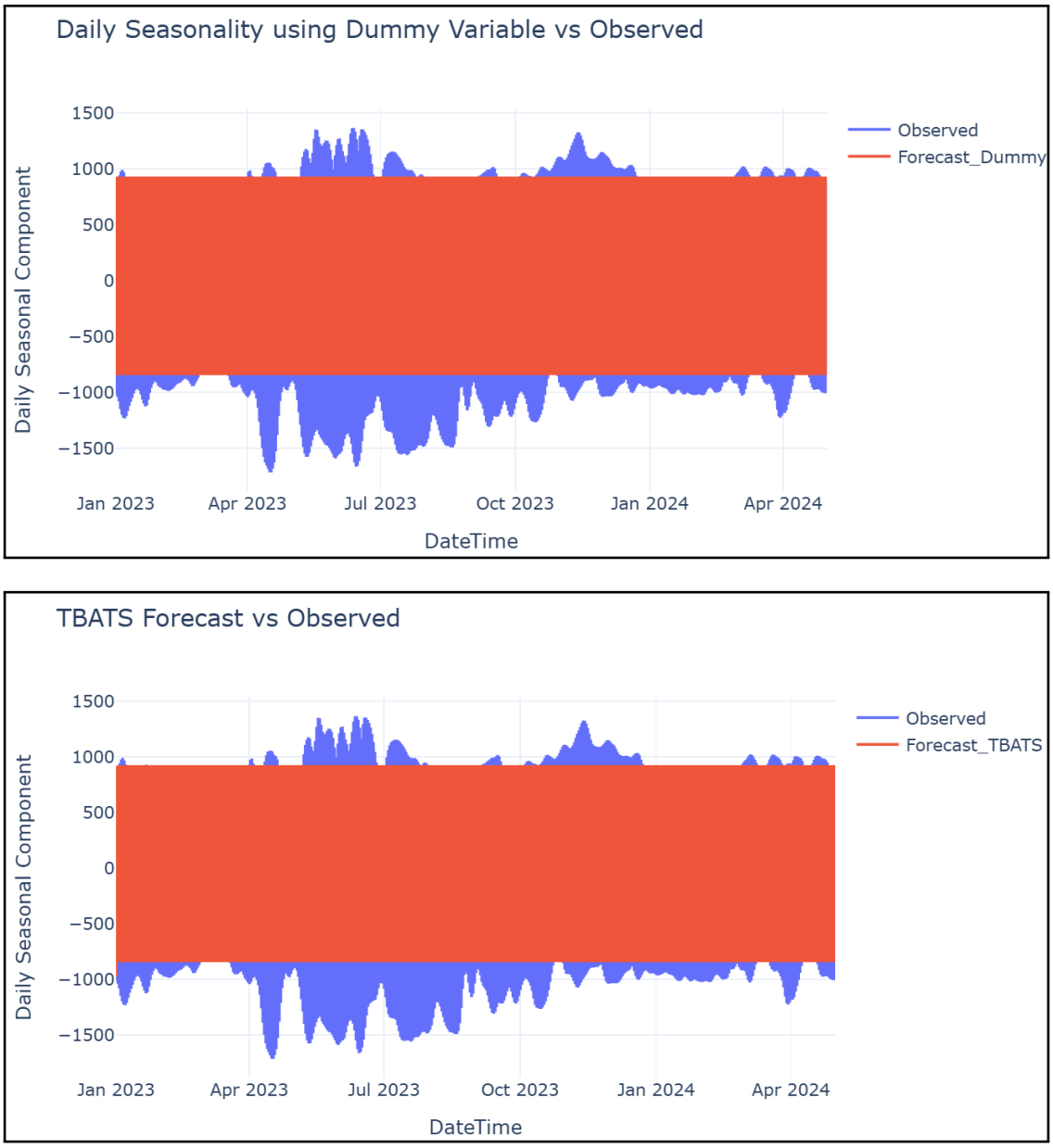


Figure 5.8: Comparison of Daily Seasonality Modeling: Dummy Variables (Top) vs TBATS (Bottom)

### 5.3.3 Modeling Yearly Seasonality: Dummy Variables vs TBATS

To capture the **yearly seasonality** (period = 8760 hours), we again considered two modeling techniques:

### 1. Dummy Variable Encoding

We created dummy variables for each **month of the year** to capture cyclical changes in electricity demand across seasons. This approach captures fixed seasonal effects such as increased demand in summer or winter.

### 2. TBATS Model

The TBATS model was again used due to its ability to model complex and long-period seasonal patterns such as yearly trends. However, TBATS may not perform well on such long seasonal cycles unless the signal is strong and regular.

The models were trained on data from **2017 to 2022** and used to predict yearly seasonality from **January 2023 to April 2024**. The table below summarizes the performance in terms of Mean Squared Error (MSE).

Table 5.2: Comparison of Yearly Seasonality Modeling Methods

| **Modeling Approach** | **MSE** | **Remarks** |
|---|---|---|
| Dummy Variables (Linear Model) | 261201.81 | Significantly lower error |
| TBATS | 1591661.18 | Underperforms on long-cycle seasona |

It is evident from the results that dummy variables performed significantly better than TBATS in capturing yearly seasonality. This suggests that the yearly seasonal component in the data is relatively stable and can be effectively modeled using fixed monthly effects rather than a flexible but more complex TBATS model.

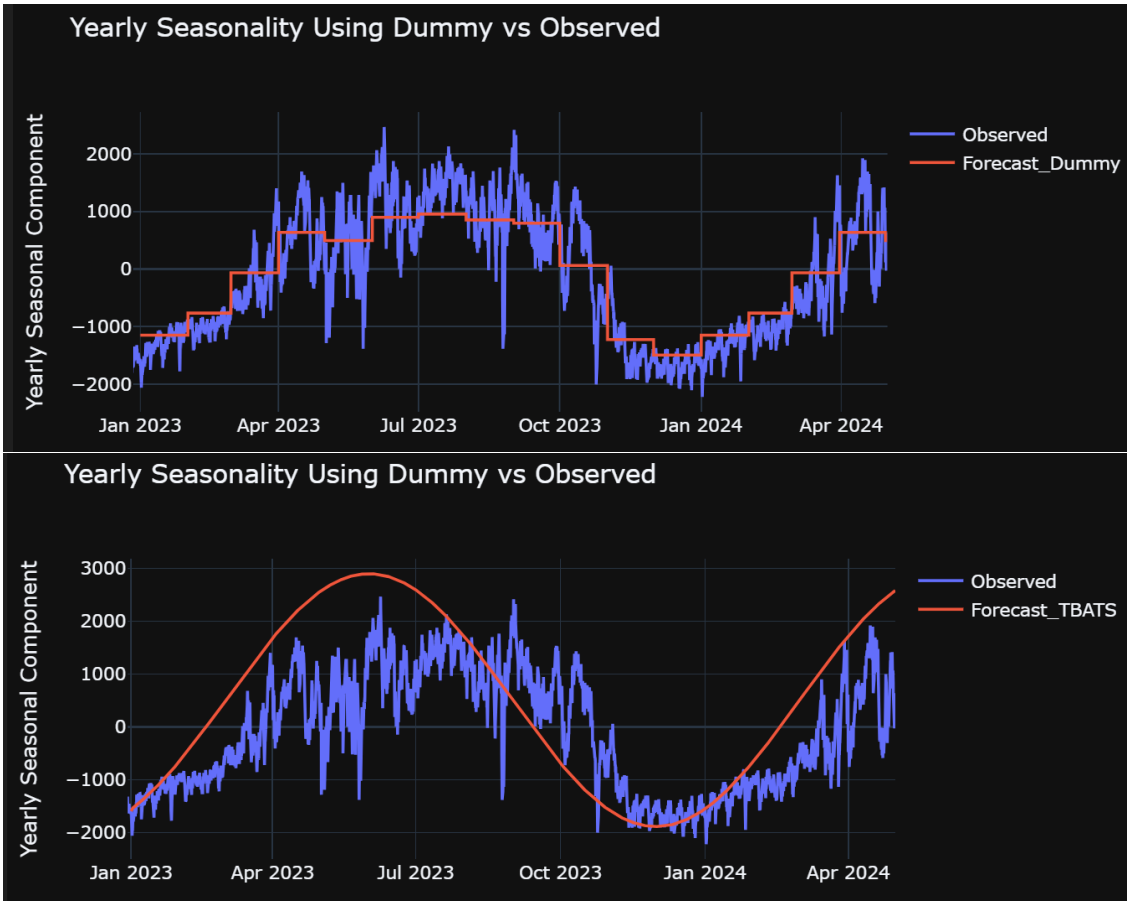


Figure 5.9: Comparison of Yearly Seasonality Modeling: Dummy Variables (Top) vs TBATS (Bottom)

## 5.4 Alternative Approach

To compare with our spline-based approach for trend estimation, we explored a simpler and more interpretable method—**linear regression**—to model the trend component extracted from the MSTL decomposition.

### 5.4.1 Methodology

The trend component obtained from the MSTL decomposition was regressed against time using a standard linear regression model. Let $T_t$ denote the trend component at time $t$. The model is specified as:

$$\hat{T}_t = \beta_0 + \beta_1 t$$

Where:

- $t$ is the time index (converted to a numerical form such as timestamp or number of hours since the start),
- $\beta_0$ is the intercept,
- $\beta_1$ is the slope representing the average rate of change over time.

The linear model was trained on data from **January 2017 to December 2022**, and forecasts were generated for the period **January 2023 to April 2024**.

### 5.4.2 Advantages and Limitations

Using data from January 2017 to December 2022, we obtained the following estimated coefficients:

$$\begin{aligned} \textbf{Linear Trend Intercept:} \quad & \beta_0 = 5548.46 \\ \textbf{Linear Trend Coefficient:} \quad & \beta_1 = 0.01865 \end{aligned}$$

### 5.4.3 Interpretation

The positive coefficient ($\beta_1$) suggests a slight upward trend in the target variable over time. Specifically, the model estimates an average increase of approximately 0.0187 units per hour.

### 5.4.4 Visualization

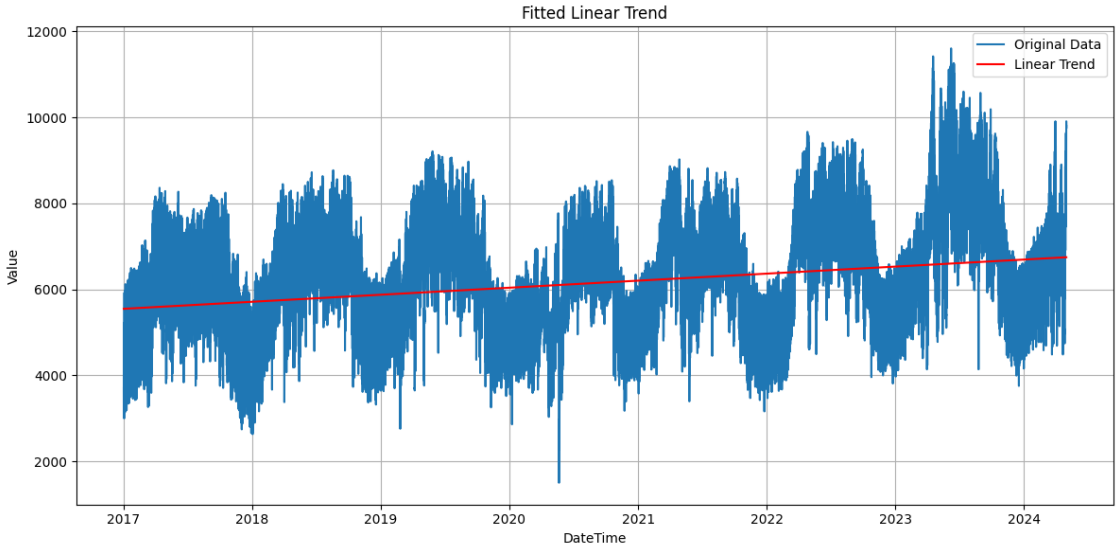


Figure 5.10: Comparison of Trend Component with Linear Regression Fit

### 5.4.5 Conclusion

The linear regression approach provides a quick benchmark for modeling the trend. However, for complex trend structures, especially in long time series with changing patterns, spline-based methods offer better flexibility and fit. This experiment helped validate that our choice of using splines was appropriate for this dataset, though linear regression remains a valuable baseline.

### 5.4.6 Modeling Daily Seasonality Using Dummy Variables

After removing the trend component (as described in the previous section using linear regression), we focused on modeling the daily seasonal structure present in the data. Since the data is hourly, a clear 24-hour periodicity was expected and confirmed by the ACF plot.

#### Methodology

To capture the daily seasonality, we created dummy variables for each hour of the day. Specifically, we introduced 24 binary variables:

$$D_0, D_1, \ldots, D_{23}$$

where $D_h = 1$ if the observation is from hour $h$ of the day, and 0 otherwise. To avoid multicollinearity (the dummy variable trap), the dummy for hour 0 ($D_0$) was dropped.

The resulting linear model is:

$$Y_t = \sum_{h=1}^{23} \gamma_h D_h + \varepsilon_t$$

Where:

- $Y_t$ is the detrended value at time $t$,
- $\gamma_h$ are the coefficients corresponding to each hour's effect relative to hour 0 (the baseline),
- $\varepsilon_t$ is the residual error.

The coefficients $\gamma_h$ thus represent the deviation from hour 0.

**Model Coefficients**

The estimated coefficients for the model were as follows:

- **Intercept (Hour 0)**: 70.95
- **Hourly Deviations ($\gamma_1$ to $\gamma_{23}$)**:

| Hour | Coefficient |
|---|---|
| 1 | -207.57 |
| 2 | -348.61 |
| 3 | -437.90 |
| 4 | -516.91 |
| 5 | -741.75 |
| 6 | -916.16 |
| 7 | -770.52 |
| 8 | -595.80 |
| 9 | -388.64 |
| 10 | -222.07 |
| 11 | -56.37 |
| 12 | 62.92 |
| 13 | 30.29 |
| 14 | 13.20 |
| 15 | -19.61 |
| 16 | -123.49 |
| 17 | 57.30 |
| 18 | 597.20 |
| 19 | 859.17 |
| 20 | 767.30 |
| 21 | 615.77 |
| 22 | 397.13 |
| 23 | 242.28 |

### Visualizations

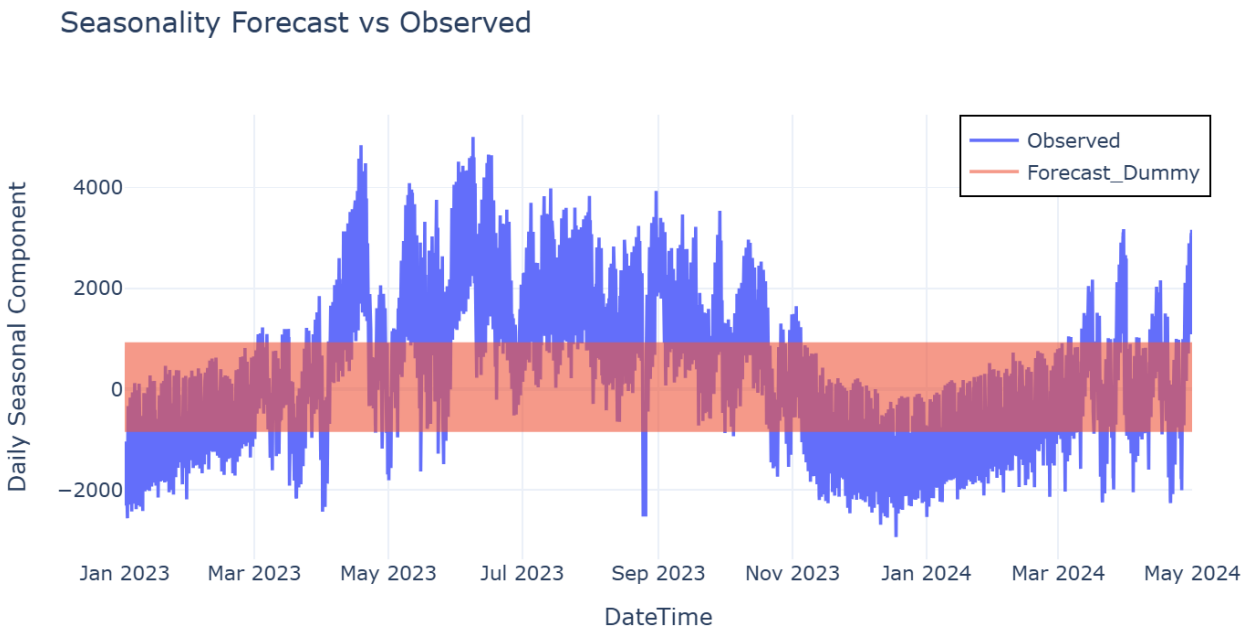


Figure 5.11: Detrended Data vs Daily Seasonality

### Prediction Period

The model was trained on data from **January 2017 to December 2022**, and predictions were generated for the period **January 2023 to April 2024**, focusing only on the daily seasonal component.

### Remarks

This model captures fixed hourly effects well. The most prominent positive deviations were observed during evening hours (18:00–21:00), while early morning hours (3:00–7:00) exhibited significant negative deviations from the baseline. While interpretable, this approach assumes a static pattern which may not capture evolving or context-dependent seasonality effectively.

## 5.4.7 Modeling Yearly Seasonality Using Dummy Variables

After removing both the trend and daily seasonality components from the original series, the residual still exhibited yearly periodic behavior, which

was evident from visual inspection and the ACF plot. To model this yearly seasonality, we used dummy variables for each month of the year.

### Methodology

Since the data is hourly, we first aggregated it to a monthly average to reduce noise and capture the long-term seasonality more effectively. We then introduced 12 dummy variables:

$$M_1, M_2, \ldots, M_{12}$$

where $M_m = 1$ if the observation belongs to the $m^{\text{th}}$ month (January to December), and 0 otherwise. To prevent multicollinearity, we dropped the dummy for January ($M_1$).

The linear model used to predict monthly averages was:

$$Y_t = \sum_{m=2}^{12} \theta_m M_m + \varepsilon_t$$

Where:

- $Y_t$ is the average value at time $t$ (monthly aggregated),
- $\theta_m$ are the coefficients representing each month's effect relative to January,
- $\varepsilon_t$ is the residual term.

The estimated model parameters are:

$$\text{Intercept } \beta_0 = -1174.54$$

$$\text{Coefficients } \beta_2 \text{ to } \beta_{12} = \begin{bmatrix} 340.47 & 1157.39 & 1885.30 & 1558.83 & 1982.87 & \\ 2113.09 & 2059.84 & 1978.09 & 1265.16 & -31.99 & -254.78 \end{bmatrix}$$

### Prediction Period

The model was trained on monthly averages from **2017 to 2022**, and predictions were made for **January 2023 to April 2024**.

### Visualization

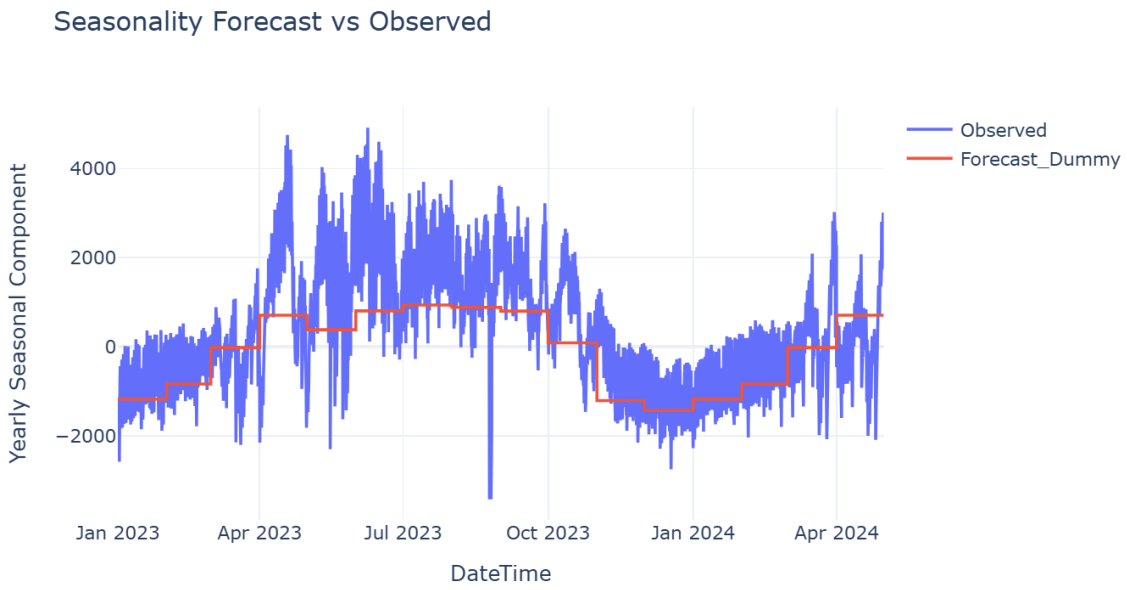


Figure 5.12: Yearly Seasonal Vs Detrended and Deseasonalized(Daily) Data

### Conclusion

Using month-wise dummy variables provides a simple and interpretable way to capture regular yearly patterns. This approach effectively captures fixed monthly effects, revealing clear yearly seasonality. The high positive coefficients for summer months indicate a peak in electricity demand, whereas the coefficients for November and December are lower. The model assumes that the seasonal effect of each month is constant over the years, which may be a limitation in the presence of evolving seasonal behavior.

## 5.4.8 Modeling Weekend Effect Using Dummy Variable

After accounting for trend, daily, and yearly seasonality, we examined whether weekends had a distinct impact on electricity demand. Based on exploratory data analysis, we introduced a binary variable to capture weekend behavior.

### Feature Engineering

A new column `is_weekend` was created as follows:

$$\texttt{is\_weekend}_t = \begin{cases} 1 & \text{if } t \text{ is a Saturday or Sunday} \\ 0 & \text{otherwise} \end{cases}$$

**Model Specification**

We modeled the residual component using a linear regression model:

$$Y_t = \beta_0 + \beta_1 \cdot \texttt{is\_weekend}_t + \varepsilon_t$$

The estimated values of the parameters were:

$$\hat{\beta}_0 = 46.23, \quad \hat{\beta}_1 = -161.81$$

**Interpretation**

The intercept $\hat{\beta}_0 = 46.23$ indicates the average residual electricity demand during the weekdays. The coefficient $\hat{\beta}_1 = -161.81$ suggests that, on average, the residual demand drops by approximately 161.81 units during the weekends. This statistically significant difference supports the inclusion of a weekend-specific adjustment in the final forecasting model.

### 5.4.9 ADF Test on Residuals from Alternate Approach

To determine whether the residuals from the alternate approach exhibit stationarity, the Augmented Dickey-Fuller (ADF) test was conducted. The null hypothesis for this test is that the series has a unit root (i.e., it is non-stationary), while the alternative hypothesis indicates stationarity.

- **ADF Test Statistic:** `-14.9493`
- **p-value:** $1.29 \times 10^{-27}$
- **Used lags:** 61
- **Number of observations:** 64186
- **Critical Values:**
  - 1% level: `-3.4305`

- 5% level: `-2.8616`
- 10% level: `-2.5668`

**Interpretation:** The test statistic (`-14.9493`) is significantly lower than all the critical values at 1%, 5%, and 10% significance levels. Additionally, the p-value ($1.29 \times 10^{-27}$) is far below typical significance thresholds. Hence, we reject the null hypothesis and conclude that the residuals from the alternate approach are **stationary**, implying the model has effectively removed trend and seasonality components.

## Unexplained Autocorrelation and Possible Exogenous Effects

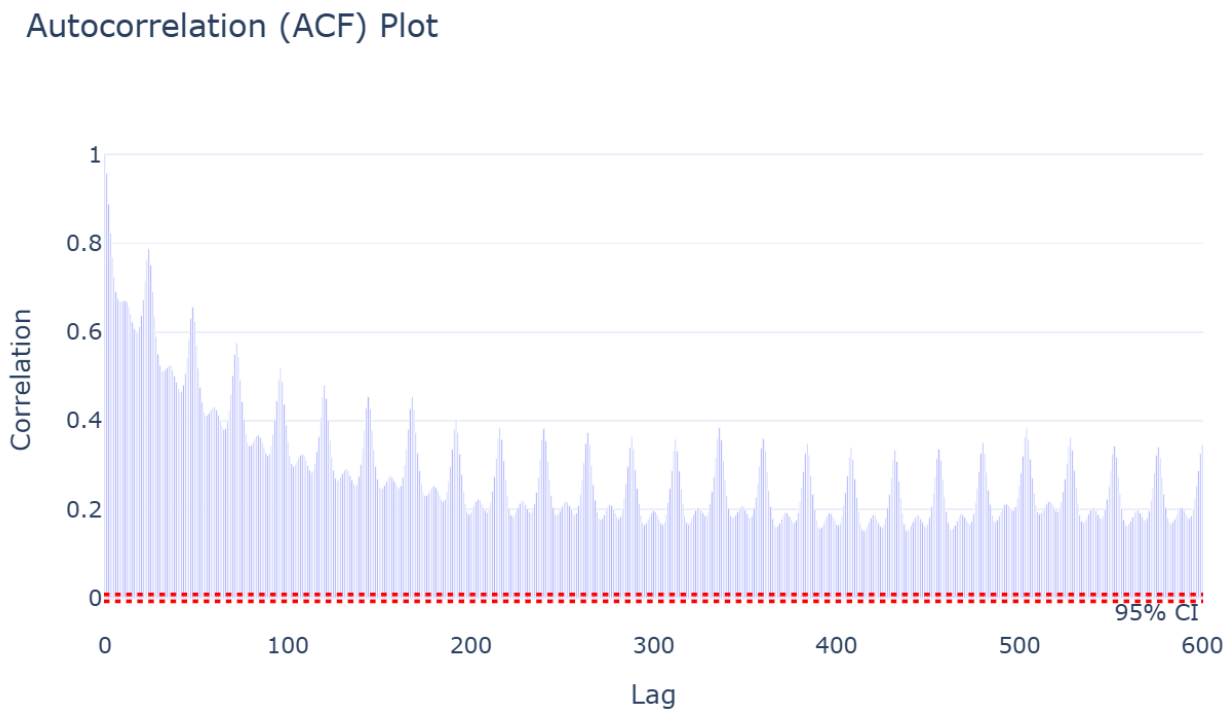


Figure 5.13: ACF Plot for the Residuals

Despite detailed decomposition using regression-based modeling of trend and seasonal components, the ACF plot of the final residuals reveals a noticeable spike at lag 24. This suggests the presence of a remaining daily cycle that is not fully accounted for by the current model.

### 5.4.10 Residual Modeling with SARIMAX

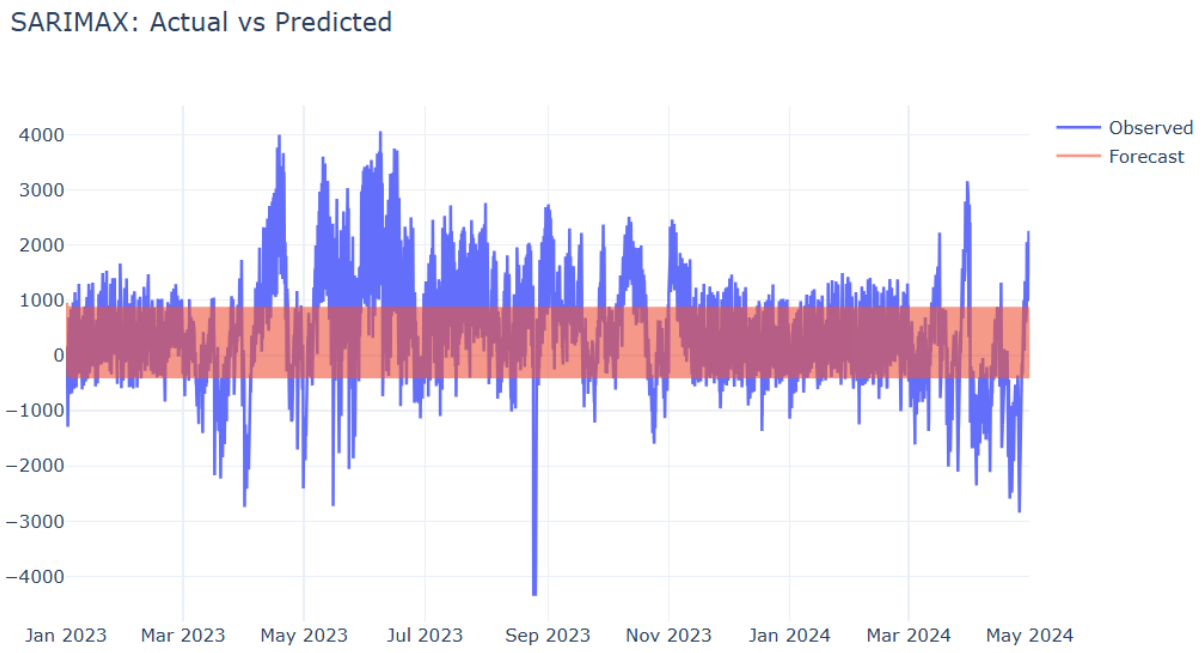


Figure 5.14: Actual Vs Predicted Sarimax Residuals

Based on the persistent autocorrelation at lag 24 observed in the residuals, we fitted a seasonal ARIMA model with exogenous structure (SARIMAX) to the residual series. The chosen specification (based on ACF/PACF inspection and AIC-guided tuning) was:

$$\text{SARIMAX}\big(p,d,q\big)\times\big(P,D,Q\big)_s \quad = \quad \text{SARIMAX}(1,0,1)\times(1,1,1)_{24},$$

i.e. a non-seasonal AR(1), MA(1) and a seasonal AR(1), seasonal MA(1) with seasonal differencing $D=1$ and seasonal period $s=24$ (hours).

**Estimated Parameters (MLE)**

| Parameter | Estimate | Std. Error | z-value | p-value |
|---|---|---|---|---|
| ar.L1 | 0.9470 | 0.001 | 842.642 | $< 0.001$ |
| ma.L1 | 0.2510 | 0.003 | 96.770 | $< 0.001$ |
| ar.S.L24 | 0.1022 | 0.004 | 27.919 | $< 0.001$ |
| ma.S.L24 | -0.8407 | 0.002 | -372.717 | $< 0.001$ |
| sigma_2 | $1.904 \times 10^4$ | 45.365 | 419.775 | $< 0.001$ |

Table 5.3: SARIMAX(1,0,1)(1,1,1)$_{24}$ — parameter estimates.

All coefficients are highly significant (p-value $< 0.001$). Notable features:

- The non-seasonal AR(1) coefficient $\phi_1 \approx 0.947$ indicates strong short-term persistence.
- The seasonal AR at lag 24 is positive but small ($\Phi_1 \approx 0.102$), while the seasonal MA at lag 24 is large and negative ($\Theta_1 \approx -0.841$), indicating substantial seasonal shock correction.
- The estimated innovation variance is $\sigma^2 \approx 1.90 \times 10^4$.

**Diagnostics and Goodness-of-Fit**

Key diagnostics from the fitted model:

- **Ljung–Box (Q) test:** $Q$-statistic $p \approx 0.72$ (fail to reject $H_0$ of no remaining autocorrelation at the tested lags).
- **Jarque–Bera:** large statistic with $p \approx 0.00 \Rightarrow$ residuals deviate from normality (heavy tails / leptokurtic).
- **Heteroskedasticity (H) test:** mild evidence of heteroskedasticity (reported $H \approx 1.47$, Prob(H) small), suggesting residual variance may not be strictly constant.
- **AIC / BIC:** model AIC/BIC reported in the fit were used during selection (reported values available in model output).

### Forecast Performance (Test Period)

After adding SARIMAX-predicted residuals to the component forecasts (trend + daily seasonality + monthly seasonality + sarimax residuals), the final forecast errors on the test set were:

- **MSE:** 1,050,565.25
- **MAE:** 735.59

### Visualization

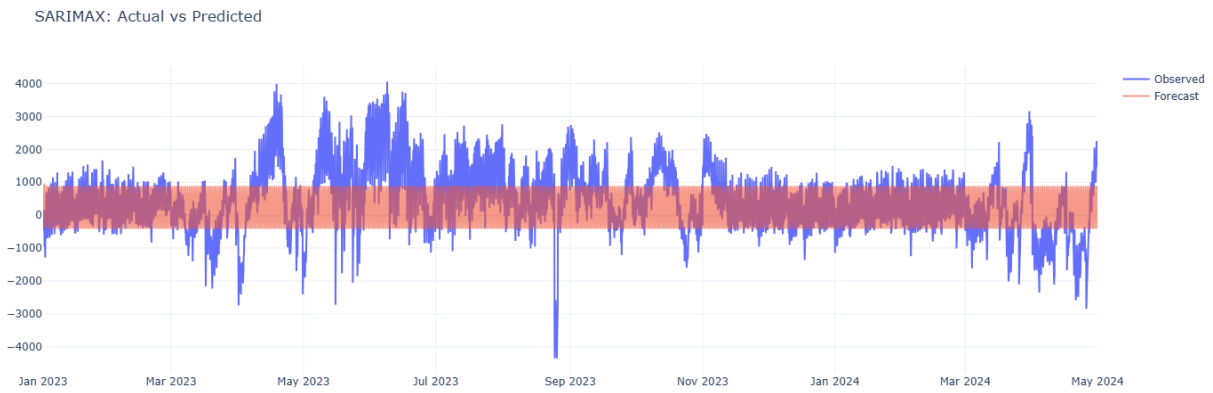


Figure 5.15: Actual Vs Predicted Hourly Electricity Demand

These metrics indicate a reduction in both MSE and MAE relative to the prior SARIMAX attempts (and relative to some earlier variants), i.e. the current fit improved numeric accuracy. However, the ACF of the combined residuals still shows spikes at lag 24 (daily), meaning some daily autocorrelation structure remains.

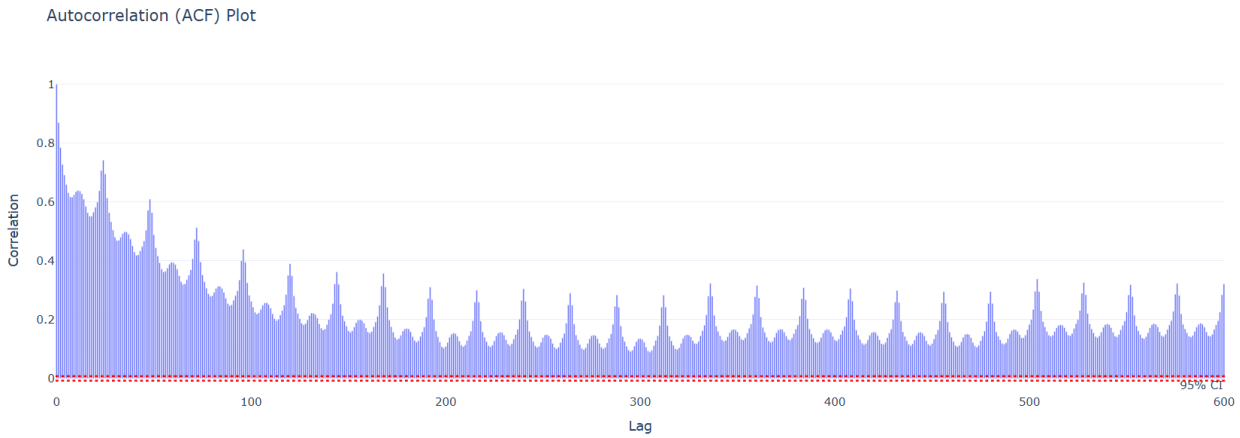


Figure 5.16: ACF Plot of Sarimax Residuals

### Interpretation and Practical Conclusion

1. **Modeling success:** The SARIMAX$(1,0,1)(1,1,1)_{24}$ model captures much of the short-term and seasonal dependence (significant AR/MA seasonal and non-seasonal terms; Ljung–Box indicates little remaining autocorrelation at tested lags).

2. **Remaining seasonal structure:** Persistent spikes at lag 24 in the post-fit ACF imply that either (a) the seasonal structure is more complex than the linear seasonal ARMA terms capture, or (b) there are *exogenous* drivers with a daily cycle that are not included in the regressors.

3. **Non-normal residuals and heteroskedasticity:** The Jarque–Bera and heteroskedasticity tests reveal heavy tails and variance heterogeneity. These affect confidence intervals and suggest possible improvements (e.g., robust errors, heavy-tailed innovations such as Student-$t$, or a GARCH-type variance model).

4. **Practical implication:** Although point-forecast accuracy improved (MSE/MAE decreased), the persistence of a lag-24 signal and non-Gaussian residuals indicate that:

    - additional exogenous covariates (e.g., temperature, scheduled industrial cycles, known operational events, calendar effects) should be considered, or

- a more flexible model class (e.g., TBATS, state-space models with time-varying parameters, regime-switching models, or models with heavy-tailed errors) may be required.

**Recommendation**

To reduce the remaining daily autocorrelation and improve reliability of forecast intervals:

- Attempt to collect and include candidate exogenous variables with daily structure (weather, temperature, demand-side interventions, scheduled maintenance, public holidays, known tariff or policy shifts).
- Experiment with alternative residual distributions (Student-$t$) or heteroskedastic models (GARCH) if volatility is time-varying.
- Consider hybrid frameworks (e.g., MSTL decomposition $\rightarrow$ component-wise models $\rightarrow$ ML model for residuals) or non-linear models that can better capture complex seasonal interactions.

In summary, the SARIMAX fit improved forecast accuracy, produced statistically significant seasonal and non-seasonal parameters, and passed basic correlation diagnostics, but the lingering lag-24 ACF spikes and non-normal residuals point to unobserved/exogenous drivers or model misspecification that merit further investigation.

## Remarks of the Alternate Approach

In the alternative modeling strategy, the trend was captured using a linear regression model. The detrended series was then used to model daily, weekly, and yearly seasonalities through dummy variables, where a level was dropped in each case to avoid multicollinearity. A weekend effect was also incorporated via a binary dummy variable. Subsequently, the residuals from this decomposition-based approach were modeled using a SARIMAX framework to capture any remaining autocorrelations and unmodeled dynamics. This hybrid approach yielded a reduced Mean Squared Error (MSE) of **1,050,565.25** and a Mean Absolute Error (MAE) of **735.59**, compared to the original 1,147,696.55 and 808.91, respectively. The results indicate that augmenting the simpler linear decomposition model with SARIMAX

over the residuals can provide notably improved accuracy while preserving interpretability and implementation ease.

## 5.5 Comparison of the MSTL and Alternate Approach

The alternate approach using linear regression with dummy variables showed a noticeable spike in the ACF plot at lag 24. This suggests that some autocorrelation remains unmodeled in that approach. The lag-24 spike typically indicates a 24-hour cyclical pattern that may not have been fully captured using simple dummy variables.

### 5.5.1 Possible Causes of Residual Autocorrelation at Lag 24

The presence of significant autocorrelation at lag 24 in the alternate approach may be due to:

- **Omitted hourly dependencies:** The current hour may depend on the same hour from the previous day (lag 24), which was not explicitly modeled.
- **External factors:** Exogenous variables such as temperature, workload, or scheduled processes may follow a daily cycle and influence the time series.
- **Nonlinear effects or interactions:** Simple linear models with additive dummy variables may fail to capture complex seasonal behavior or dependencies.

### 5.5.2 Remarks

The MSTL decomposition provided better modeling of seasonal patterns as evident from the white-noise nature of the residuals. The alternate dummy-variable-based approach may benefit from the inclusion of additional features or a more flexible model structure to capture autocorrelated patterns.

# Chapter 6

# Conclusion

This study evaluated three different modeling strategies for forecasting the time series data: (1) MSTL decomposition with dummy variables, (2) MSTL decomposition using TBATS for seasonality modeling, and (3) an alternate approach using linear models. Additionally, the alternate approach was further enhanced by fitting a SARIMAX model on its residuals to capture remaining autocorrelations and unmodeled seasonal effects. The performance of each method was assessed using the Mean Squared Error (MSE) and Mean Absolute Error (MAE) on the final forecasts. The results are summarized below:

- **MSTL with Dummy Variables:**
  - MSE: `1,200,957.40`
  - MAE: `834.03`
- **MSTL with TBATS:**
  - MSE: `1,764,012.47`
  - MAE: `1037.78`
- **Alternate Approach (Linear Decomposition with Dummies):**
  - MSE: `1,147,696.55`
  - MAE: `808.91`
- **Alternate Approach + SARIMAX on Residuals:**

- MSE: 1,050,565.25
- MAE: 735.59

From the metrics, it is evident that augmenting the **alternate approach** with SARIMAX modeling over its residuals produced the most accurate forecasts, with the lowest MSE and MAE values. While the MSTL model with dummy variables and the plain alternate approach performed well, the SARIMAX-enhanced alternate approach further reduced forecast errors, suggesting it effectively captured additional patterns—possibly including unobserved external influences—beyond what the initial linear decomposition and dummy variables could model. The MSTL model utilizing TBATS for seasonality adjustment showed the weakest performance among the four. Thus, the SARIMAX-augmented alternate modeling strategy proved to be the most effective in capturing the underlying patterns in the data and producing reliable forecasts.